Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization

Vikas Peddu[1], Ryan C. Shean[1], Hong Xie[1], Lasata Shrestha[1], Garrett A. Perchetti[1], Samuel S. Minot[2], Pavitra Roychoudhury[1,2], Meei-Li Huang[1], Arun Nalla[1], Shriya B. Reddy[3], Quynh Phung[1], Adam Reinhardt[1], Keith R. Jerome[1,2#], Alexander L. Greninger[1,2,#]

[1]Department of Laboratory Medicine, University of Washington, Seattle, WA, USA

[2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[3]Boston University School of Medicine, Boston, MA, USA

[#]Corresponding authors/contributed equally:

Alexander L. Greninger: agrening@uw.edu (Phone: (415) 439-3448, Fax: (206) 598-6189),

Keith R. Jerome: kjerome@fredhutch.org (Phone: (206) 667-6793, Fax: (206) 598-6189)

Abbreviations:
Metagenomic next-generation sequencing (mNGS)
Human parainfluenza virus type 3 (HPIV3)
Reads per million (RPM)
Sever acute respiratory syndrome-related coronavirus (SARS-related coronavirus)

Abstract

Background: More than two months separated the initial description of SARS-CoV-2 and discovery of its widespread dissemination in the United States. Despite this lengthy interval, implementation of specific quantitative reverse transcription (qRT)-PCR-based SARS-CoV-2 tests in the US has been slow, and testing is still not widely available. Metagenomic sequencing offers the promise of unbiased detection of emerging pathogens, without requiring prior knowledge of the identity of the responsible agent or its genomic sequence.

Methods: To evaluate metagenomic approaches in the context of the current SARS-CoV-2 epidemic, laboratory-confirmed positive and negative samples from Seattle, Washington were evaluated by metagenomic sequencing, with comparison to a 2019 reference genomic database created before the emergence of SARS-CoV-2.

Results: Within 36 hours our results showed clear identification of a novel human *Betacoronavirus*, closely related to known Betacoronaviruses of bats, in laboratory-proven cases of SARS-CoV-2. A subset of samples also showed superinfection or colonization with human parainfluenza virus 3 or *Moraxella* species, highlighting the need to test directly for SARS-CoV-2 as opposed to ruling out an infection using a viral respiratory panel. Samples negative for SARS-CoV-2 by RT-PCR were also negative by metagenomic analysis, and positive for Rhinovirus A and C. Unlike targeted SARS-CoV-2 qRT-PCR testing, metagenomic analysis of these SARS-CoV-2 negative samples identified candidate etiological agents for the patients' respiratory symptoms.

Conclusion: Taken together, these results demonstrate the value of metagenomic analysis in the monitoring and response to this and future viral pandemics.

Introduction

On January 20, 2020, less than one month after the initial reports of a series of viral pneumonia cases in Wuhan, China, the first case of infection with the novel SARS-CoV-2 was confirmed in the United States (1). Rapid person-to-person transmission has resulted in 614,482 total cases and 27,085 deaths within the United States as of April 15, 2020 (2). Epidemiological analyses have shown increased mortality risk in elderly patients above 65 years age, especially with underlying comorbidities (3-4). Reported clinical complications that develop include sepsis in 59% of cases and acute respiratory distress syndrome in 17-29% of cases, often progressing to require mechanical ventilation (5-7).

For rapidly emerging infectious diseases, metagenomic next-generation sequencing (mNGS) offers an opportunity to both recover whole viral genomes for epidemiological purposes and to agnostically determine co-infections that may be associated with increased morbidity and mortality in emerging infectious diseases (8). Here, we evaluated the performance of metagenomic sequencing on eight samples sent for SARS-CoV-2 diagnostic testing. mNGS was performed in under 36 hours from sample collection to analysis, and the results were confirmed using validated qRT-PCR based methods.

Methods

Sample collection and qRT-PCR detection of SARS-CoV-2

Eight nasopharyngeal swabs in viral transport medium were sent to the University of Washington Clinical Virology laboratory for diagnostic or confirmatory

testing. qRT-PCR was performed using a modified protocol of the World Health Organization's assay for qualitative diagnostics of COVID-19 targeting SARS-CoV-2 RNA. Two hundred µL of the original sample, a Copan 403C swab in 3 mL of viral transport medium, was extracted using the Roche MagNA Pure 96 Viral NA Small Volume Kit and eluted into a final volume of 50 µL. Samples were processed and run with EXO as a spiked internal control at a concentration of 1000 copies per µL (9). Five µL of the RNA was used as input for a 40 cycle 25 µL one-step qRT-PCR reaction on an AgPath-ID system on an ABI 7500 Real-Time PCR instrument. This assay targets both the E-gene and RdRp gene. Positive results were reported when both targets were amplified. Negative results were reported when neither target was amplified.

RNA-sequencing of positive and negative samples

Eight unique patient samples consisting of six positive and two negative cases of suspected SARS-CoV-2 were sequenced using RNA extracted for a qRT-PCR diagnostic assay. In parallel we created mNGS sequencing libraries using a previously published protocol using ds-cDNA synthesis, followed by Nextera XT tagmentation and 20 cycles of PCR amplification (10). These libraries were sequenced on an Illumina MiSeq using a 1x185 run with the MiSeq Reagent Kit v3 (150-cycle).

Metagenomic analysis and interpretation of sequencing data

Raw FASTQ files were analyzed with the freely available and open source metagenomics pipeline developed at UW virology, CLOMP (11). First, raw reads were adapter trimmed and

quality filtered using Trimmomatic v0.38 (13). All reads longer than 65 base pairs with a minimum phred quality score of 30 were kept for further analysis. Entropy based low-complexity filtering was done using BBTools' BBduk to remove any reads with an entropy score lower than 0.7 using a sliding window of 50 basepairs at a kmer size of 4 (14).

Alignment of the remaining reads to HG38 was done using Bowtie2, and any aligned reads were removed at this step (15). The human-depleted reads were aligned to the Genbank nucleotide (NT) database downloaded in February 2019 using the SNAP alignment algorithm (16). Certain entries were manually removed from the build due to the presence of adapter sequence. Due to computational limitations, the NT reference database was split into 14 chunks. Therefore, each read was aligned separately to each of the 14 chunks producing multiple potential assignments for each read. These alignments were combined before read classification that follows after.

Final taxonomical classification of reads was performed by assigning each read to the most specific National Center for Biotechnology Information taxonomy ID appearing in all assignments (assignments to "other sequences", "artificial sequences", or "environmental samples" were ignored). Reads per million (RPM) calculations and inter-sample comparisons were performed using the RPM_summary.r script (11). Output from the pipeline was visualized using the Pavian metagenomics data explorer and interpreted by a bioinformatician as well as two board-certified pathologists, who were blinded to clinical information on the samples prior to interpretation (Table 1).

Analysis of unassigned reads

Reads that neither aligned to HG38 nor the NT database were re-trimmed using Trimmomatic (13). Mitochondrial sequences were depleted prior to assembly reads by alignment to the human mitochondrial genome (MN540528.1) using Bowtie2 with default options (15). Aligned reads were removed using Samtools (17), and then converted back to FASTQ format using Samtools fastq. BBTools bbfakereads.sh was used to split the single end sequence into a pseudo-paired end sequence for assembly with metaSPAdes (14, 18).

Generation of the phylogenetic tree

All available SARS-CoV-2 sequences from the Global Initiative on Sharing All Influenza Data were downloaded on 3/18/2020, consisting of 806 unique samples. Sequences with more than 5% N content were manually removed. Genome alignment was done using MAFFT with the default settings. Phylogenetic trees were built using RAxML using the GTRCATI model with 1000 bootstrap replicates.

Results

Successful detection and recovery of SARS-CoV-2 genomes

Any taxa with an RPM < 10 were filtered out in order to exclude misclassified reads, possible water contaminants, and nasal flora. Independent blinded analysis by both a bioinformatician and board-certified pathologist arrived at concordant interpretation of the results described.

Despite our reference database not containing any SARS-CoV-2 genomes, the six samples that were positive for SARS-CoV-2 by qRT-PCR had reads classified to

"Severe acute respiratory syndrome-related coronavirus" species (SARS-related coronavirus) (National Center for Biotechnology Information TaxID 694009) with a median RPM of 1171 [interquartile range 469, 2904]. SARS-CoV-2 reads were also classified to the Rhinolophus bat coronavirus BtCoV/4991, another species in the family *Coronaviridae*, for all six positive samples with a median RPM of 650 [interquartile range 244, 1819]. From the unassigned reads, we were able to assemble six full SARS-CoV-2 genomes, all with greater than 97% coverage relative to the reference SARS-CoV-2 reference strain (NC_045512.2). With as few as 941,164 reads, we were able to assemble at least 97.9% of the SARS-CoV-2 genome from a sample (WA8-UW5) with C_T = 24.8 (Table 2). We were able to similarly detect and assemble SARS-CoV-2 with a C_T of 29.5. As expected, our approach scaled with C_T (R^2= 0.80) (Table 2, Figure 2).

Evidence of a SARS-CoV-2 and HPIV3 coinfection

Sample WA6-UW3 showed substantial evidence of HPIV3 infection with an RPM of 4002 consisting of 4,027 unique reads. Reads from this sample aligning to HPIV3 were also successfully *de novo* assembled using the Geneious 9.1.8 assembler (19). From this assembly we were able to reproduce the full HIPV3 genome with a mean depth of coverage of 66.4.

SARS-CoV-2 negative samples contained Rhinovirus spp.

The two SARS-CoV-2 negative samples, SC5683 and SC5698, contained reads classifying to rhinovirus species A and C respectively. SC5683 contained reads

classifying to both rhinovirus A71 (RPM = 1,592), as well as human rhinovirus spp. (RPM = 19,061). SC5698 contained reads classifying only to rhinovirus C3 (RPM = 454) (Figure 3).

Bacterial reads were classified in addition to viral reads

Common skin flora *Cutibacterium acnes* were present to some degree in nearly all samples with a median RPM of 310.0 [interquartile range 205, 573]. *M. catarrhalis* was classified in three out of the six samples with SARS-CoV-2 (WA6-UW3, WA9-UW6, WA8-UW5), and none of the negative SARS-CoV-2 samples, with RPMs of 11, 2,305, and 5,172 (Figure 3). Of note, WA9-UW6 and WA8-UW5 had an approximately 100x higher RPM for *M. catarrhalis* than did WA6-UW3 (Figure 3). A recently prepared non-template control sequencing library yielded no bacterial reads other than *Cutibacterium acnes* (RPM 41996).

Confirmation of SARS-CoV-2 infection by qRT-PCR detection of the RNA-dependent RNA Polymerase gene

Out of the eight total samples, the six with SARS-CoV-2 detected by metagenomic sequencing had C_Ts of below 30 for both the *E* and *RdRp* genes by qRT-PCR for SARS-CoV-2. In contrast, the two samples negative for SARS-CoV-2 by metagenomic sequencing, SC5683 and SC5698, had no amplification of either gene (Table 2). The EXO internal control was successfully amplified in all tested samples.

Phylogeny reveals clustering within two distinct clades

Phylogenetic analysis revealed that the six SARS-CoV-2 sequences found cluster within two clades representing the Washington state and European outbreaks. WA3-UW1, a traveler from Korea to Washington state, was the only sequence to cluster in the European clade. All genomes were over 99.5% identical by nucleotide relative to the reference strain (NC_045512.1). WA3-UW1 contained 3 amino acid mutations in the ORF1ab gene (W816R, V1858L, V3695L), and one in the ORF3a gene (G8715V). The remaining samples had amino acid mutations in the ORF1ab gene (P2928S, P5916L, Y5953C), and one mutation in the ORF8 gene (Y9382H).

Discussion

Using mNGS, we were able to successfully detect SARS-CoV-2 in six out of six positive samples, which were also confirmed by qRT-PCR. In addition, we were able to recover nearly full SARS-CoV-2 genomes from taxonomically unassigned reads (20). The total time required for this testing was approximately 36 hours from receiving the sample to taxonomical assessment. Such rapid turnaround could prove invaluable in the future when presented with an unknown infectious agent.

The six SARS-CoV-2 sequences we present here represent two distinct clades from the pandemic: One European and one from Washington state. Sample WA3-UW1, the only sample to cluster within the European clade, was derived from a traveler from South Korea. This sample diverged early within the clade and seems to be the terminal isolate within the United States. All other samples clustered with others from Washington state and are representative of the larger Washington State outbreak.

A consequence of our reference database having been built from the 2019 Genbank NT database is that it does not contain any SARS-CoV-2 sequences. Despite this, reads with sequence homology to SARS-CoV were able to classify SARS-CoV-2 reads to the taxa “Severe acute respiratory syndrome-related coronavirus” (National Center for Biotechnology Information Taxid 694009). This was confirmed with assembly of unassigned reads, from which we were able to retrieve nearly the whole SARS-CoV-2 genome for all positive samples. We also demonstrate that with as few as 5 million reads we can de-novo assemble a full SARS-CoV-2 genome from a sample with a C_T as high as 29.5 (Figure 2). Of note, the number of SARS-CoV-2 reads is driven not only by viral load, but also the number of reads going to bacterial or human sequence (online Supplemental Table 1). Additionally, reference genome length is not taken into account in this implementation of our pipeline. As a result, a bias is introduced as RPM will scale with the abundance of the organism, as well as increase linearly with the length of its genome.

In addition to being positive for SARS-CoV-2, samples WA6-UW3, WA8-UW5, and WA9-UW6 also showed positive metagenomic results for *M. catarrhalis*, a gram-negative diplococcus that colonizes the nares of up to 75% of children, but only 0-4% of adults (21-22). The RPM of *M. catarrhalis* from WA8-UW5 and WA9-UW6 were 100x higher than that of WA6-UW3. These results further illustrate the ability of mNGS to detect bacterial infections and/or colonizations on a patient-by-patient basis.

WA6-UW3 was the only sample out of the cohort to be positive for two viruses: SARS-CoV-2 and HPIV3. To date, this has only been demonstrated once before in the context of SARS-CoV-2 (23). At the time of mNGS analysis, Centers for Disease

Control and Prevention criteria for SARS-CoV-2 testing in persons lacking epidemiological linkages to the known infections specifically required severe lower respiratory tract infection that was negative for known pathogens. These findings demonstrate that detection of a known respiratory pathogen does not rule out the possibility of co-infection with SARS-CoV-2, and highlights the need to test directly for SARS-CoV-2 in order to rule out infection. This has important epidemiological implications due to the ongoing shortage of SARS-CoV-2 testing kits, since non-SARS-CoV2 respiratory virus testing may lead to assumed negative SARS-CoV-2 status in coinfected patients. Based on this finding we believe those individuals who have presented with symptoms of lower respiratory tract infection and had SARS-CoV-2 infection ruled out based on a positive viral respiratory panel test should be retested specifically for SARS-CoV-2 if testing is available.

In summary, we show that metagenomic sequencing represents a powerful tool for pathogen identification during emerging pandemics. We successfully identified and recovered SARS-CoV-2 genomes using unbiased methods and showed evidence of both bacterial and viral coinfections. In the future this approach could greatly increase the speed at which complex coinfections are correctly diagnosed and managed, potentially saving lives in the process.

Declarations

The authors declare they have no competing interests.

Data availability

Raw reads for SARS-CoV-2 positive samples as detected by a directed alignment against the SARS-CoV-2 reference genome (NC_045512.2) are available on the SRA under Bioproject PRJNA610428. Assembled SARS-CoV-2 genomes are available on Genbank under the following accessions: MT163716.1, MT163717.1, MT163718.1, MT163719.1, MT163720.1, MT163721.1.

Figure Legends

Figure 1. Flowchart of experimental design and pathogens detected in the samples (*Moraxella catarrhalis*, Human parainfluenza virus 3, Rhinovirus A, Rhinovirus C).

Figure 2. Relationship of SARS-related coronavirus reads per million to viral load as estimated by CT of SARS-CoV-2 specific qRT-PCR.

Figure 3. Organisms identified by metagenomic analysis. A-E) Reads per million of pathogenic organisms as determined. F) Reads per million of *Cutibacterium acnes*, common in skin flora, detected in all samples.

Table 1. Reads per million values for all samples for HPIV3, *M. catarrhalis, and* human rhinoviruses A and C.

TaxID (NCBI)	**Classification**	**WA6-UW3**	**WA7-UW4**	**WA4-UW2**	**SC5683**	**WA3-UW1**	**SC5698**	**WA9-UW6**	**WA UW**
694009	SARS-related coronavirus	4423	3474	1149	0	243	0	24	119
11216	Human Parainfluenza virus 3	4002	0	0	0	0	0	1	0
480	M. catarrhalis	17	0	0	0	0	0	687	565
147711	Human rhinovirus A	0	0	0	1592	0	0	0	0
463676	Human rhinovirus C	0	0	0	0	0	454	0	0

Table 2. Number of reads on sample, percent genome assembled from unassigned reads, reads per million (RPM) values, and RdRp gene cycle threshold (C_T) for the eight sequenced samples. Samples were reported as not detected (NDT) if there was no amplification.

Sample	Total reads on sample	Percent of SARS-CoV-2 genome assembled	SARS-related coronavirus RPM	RdRp gene C_T
WA6-UW3	1,927,886	99.8	4423	20.7
WA9-UW6	5,756,216	99.0	24	29.5
WA7-UW4	1,770,266	98.7	3474	21.7
WA3-UW1	18,419,147	98.6	243	22.9
WA8-UW5	941,164	97.9	1194	24.8
WA4-UW2	2,713,586	97.6	1149	22.8
SC5683	1,728,462	0	0	NDT
SC5698	1,013,934	0	0	NDT

RdRp, RNA-dependent RNA polymerase.

References

1. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med. 2020;382:929–36.

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis [Internet]. 2020 [cited 2020 Mar 8]; Available from: http://www.sciencedirect.com/science/article/pii/S1473309920301201

3. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. American Medical Association; 2020;323:1239–42.

4. Chen Y, Li L. SARS-CoV-2: virus dynamics and host response. Lancet Infect Dis [Internet]. Elsevier; 2020 [cited 2020 Apr 13];0. Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30235-8/abstract

5. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med [Internet]. 2020 [cited 2020 Mar 8]; Available from: https://linkinghub.elsevier.com/retrieve/pii/S2213260020300795

6. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. Elsevier; 2020;395:507–13.

7. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395:497–506.

8. Li Y, Deng X, Hu F, Wang J, Liu Y, Huang H, et al. Metagenomic analysis identified co-infection with human rhinovirus C and bocavirus 1 in an adult suffering from severe pneumonia. J Infect. 2018;76:311–3.

9. Kuypers J, Wright N, Ferrenberg J, Huang M-L, Cent A, Corey L, et al. Comparison of Real-Time PCR Assays with Fluorescent-Antibody Assays for Diagnosis of Respiratory Virus Infections in Children. J Clin Microbiol. 2006;44:2382–8.

10. Greninger AL, Zerr DM, Qin X, Adler AL, Sampoleo R, Kuypers JM, et al. Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. J Clin Microbiol. 2017;55:177–82.

11. greninger-lab/SARS-CoV-2_CC_publication [Internet]. greninger-lab; 2020 [cited 2020 Mar 27]. Available from: https://github.com/greninger-lab/SARS-CoV-2_CC_publication

12. Breitwieser FP, Salzberg SL. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. Schwartz R, editor. Bioinformatics [Internet]. 2019 [cited 2020 Mar 7]; Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz715/5573755

13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.

14. BBMap [Internet]. SourceForge. [cited 2020 Mar 9]. Available from: https://sourceforge.net/projects/bbmap/

15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

16. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and More Accurate Sequence Alignment with SNAP. :10.

17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

18. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.

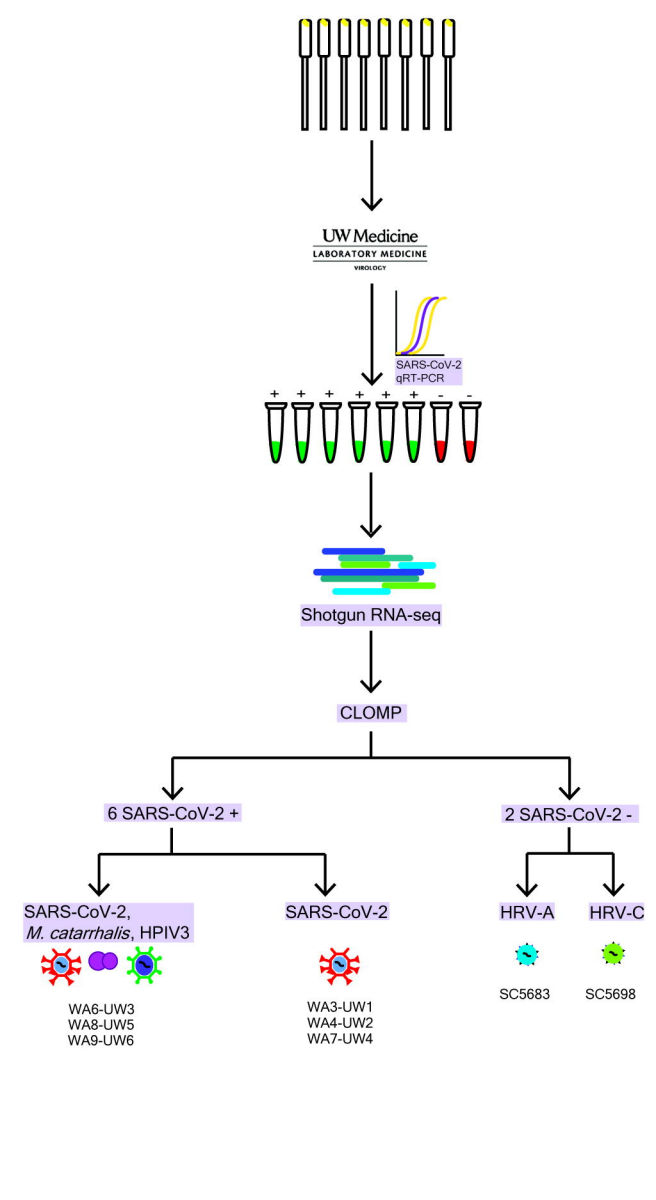
19. Geneious | Bioinformatics Software for Sequence Data Analysis [Internet]. Geneious. [cited 2020 Mar 9]. Available from: https://www.geneious.com/

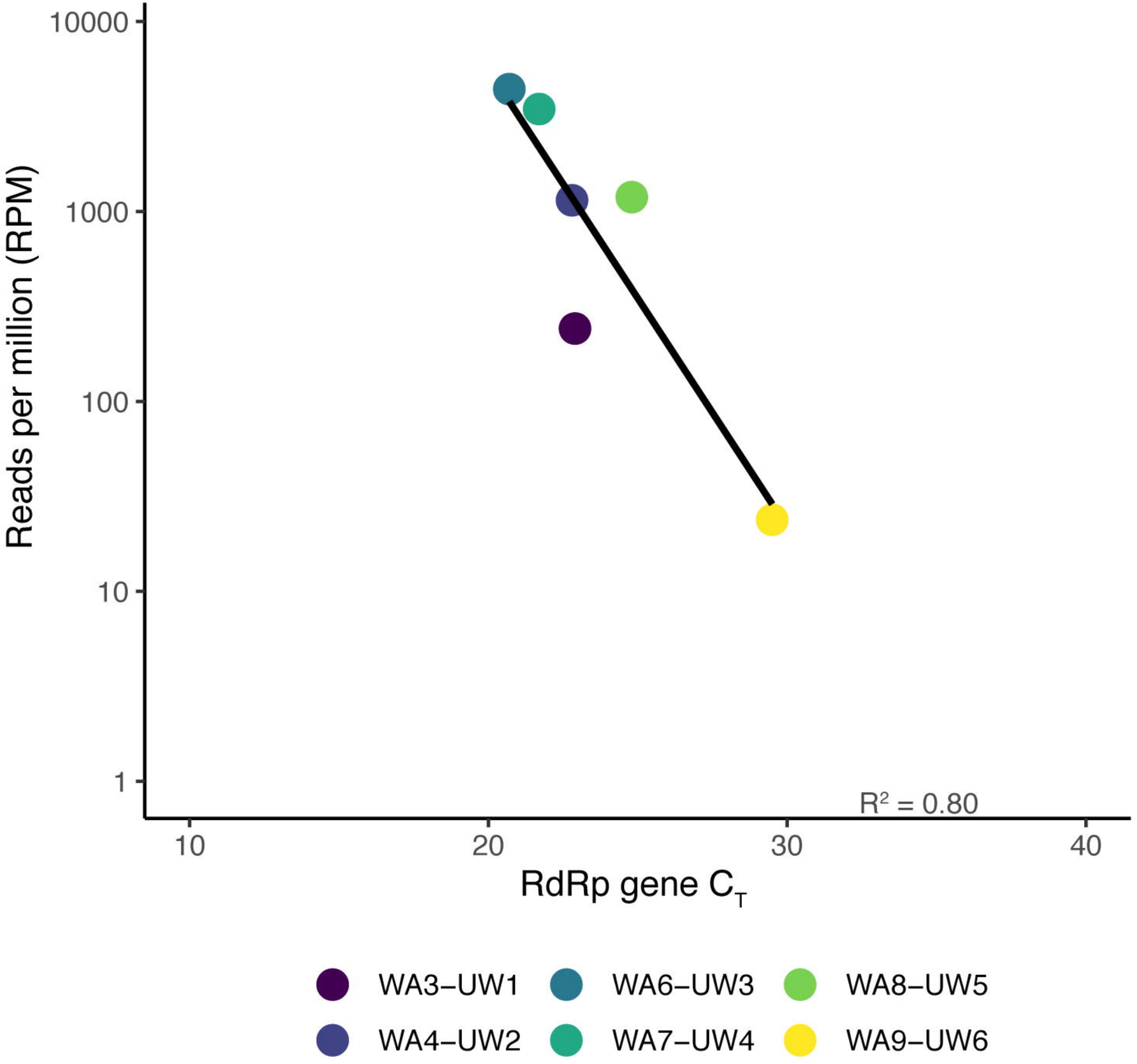
20. Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, et al. A Metagenomic Analysis of Pandemic Influenza A (2009 H1N1) Infection in Patients from North America. PLoS ONE [Internet]. 2010 [cited 2020 Mar 18];5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2956640/

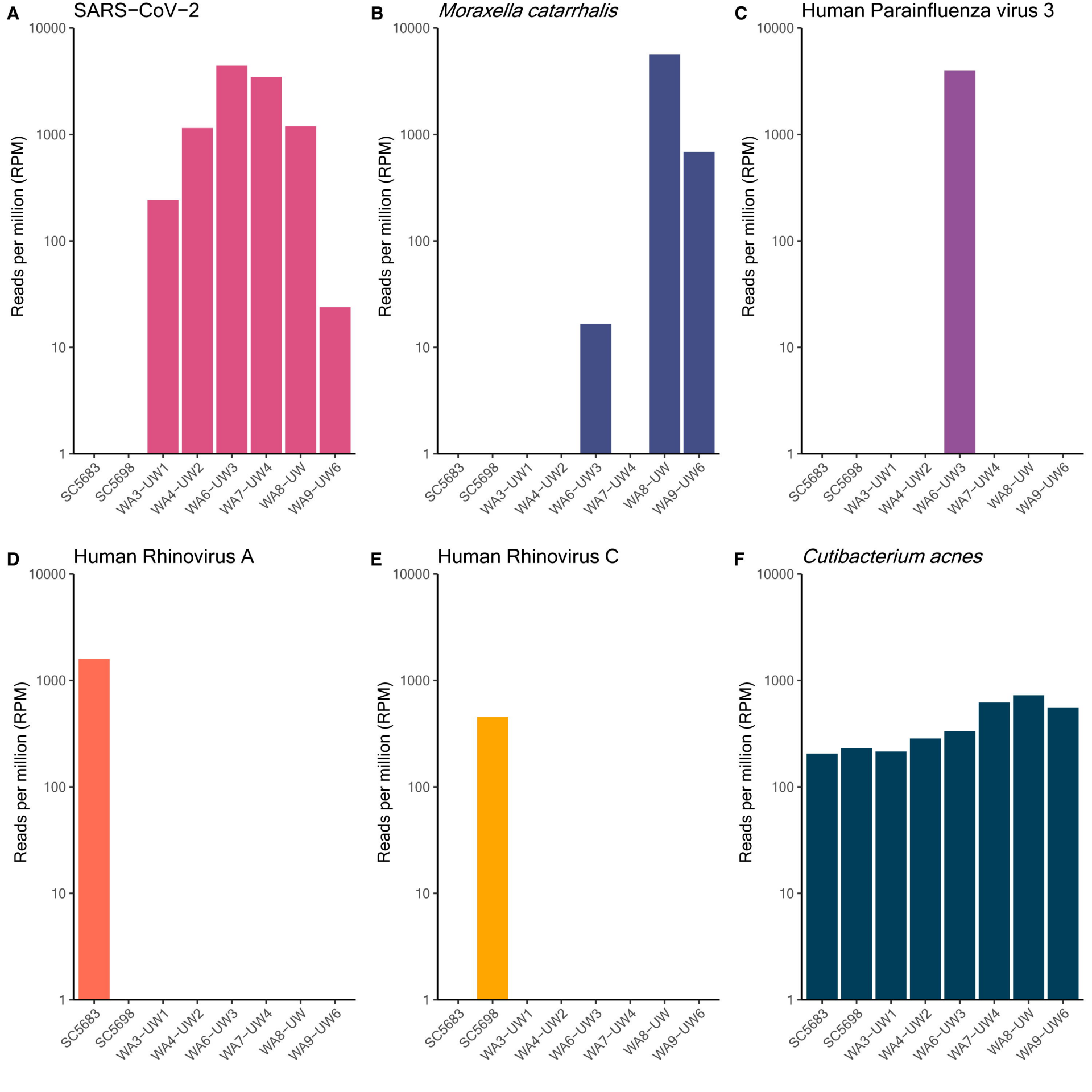
21. Verduin CM, Hol C, Fleer A, van Dijk H, van Belkum A. Moraxella catarrhalis: from Emerging to Established Pathogen. Clin Microbiol Rev. 2002;15:125–44.

22. Brook I. Microbiology of Sinusitis. Proc Am Thorac Soc. 2011;8:90–100.

23. Wu X, Cai Y, Huang X, Yu X, Zhao L, Wang F, et al. Co-infection with SARS-CoV-2 and Influenza A Virus in Patient with Pneumonia, China. Emerg Infect Dis [Internet]. 2020 [cited 2020 Mar 16];26. Available from: http://wwwnc.cdc.gov/eid/article/26/6/20-0299_article.htm


UW Medicine
LABORATORY MEDICINE
VIROLOGY
SARS-CoV-2
qRT-PCR
+ + + + + + - -
Shotgun RNA-seq
CLOMP
6 SARS-CoV-2 +
2 SARS-CoV-2 -
SARS-CoV-2,
M. catarrhalis, HPIV3
SARS-CoV-2
HRV-A
HRV-C
WA6-UW3
WA8-UW5
WA9-UW6
WA3-UW1
WA4-UW2
WA7-UW4
SC5683
SC5698


10000
1000
100
10
1
Reads per million (RPM)
10
20
30
40
RdRp gene C_T
R^2 = 0.80
WA3–UW1
WA4–UW2
WA6–UW3
WA7–UW4
WA8–UW5
WA9–UW6


A
SARS−CoV−2
B
Moraxella catarrhalis
C
Human Parainfluenza virus 3
D
Human Rhinovirus A
E
Human Rhinovirus C
F
Cutibacterium acnes
Reads per million (RPM)
10000
1000
100
10
1
SC5683
SC5698
WA3−UW1
WA4−UW2
WA6−UW3
WA7−UW4
WA8−UW
WA9−UW6